

Valid inference from non-ignorable network sampling designs

Simon Lunagomez, Edoardo M. Airolidi

Department of Statistics

Harvard University, Cambridge, MA 02138, USA

Abstract

Consider individuals interacting in social network and a response that can be measured on each individual. We are interested in making inferences on a population quantity that is a function of both the response and the social interactions. In this paper, working within Rubin's inferential framework, we introduce a new notion of non-ignorable sampling design for the case of missing covariates. This notion is the key element for developing valid inferences in applications to epidemiology and healthcare in which hard-to-reach populations are sampled using link-tracing designs, including respondent-driven sampling, that carry information about the quantify of interest.

Keywords: Bayesian inference; finite and infinite population estimands; frequentist coverage; model averaging; respondent driven sampling; HIV study.

Contents

1	Introduction	1
2	Theory and Definitions	4
2.1	The Concept of Ignorability in the Context of Social Networks	4
2.2	Respondent-Driven Sampling	5
3	Statistical Methodology	6
3.1	Modeling framework	6
3.2	Posterior inference and estimation	7
3.3	Model Specification	9
4	Markov Chain Monte Carlo Algorithm	10
4.1	Update ψ	11
4.2	Update ζ	12
4.3	Update Y_{AUG}	12
4.4	Update \mathcal{G}_{AUG}	13
4.5	Proposals	14
5	Results	15
5.1	Simulation Study	15
5.2	MCMC Performance	16
5.3	When the Prior on \mathcal{G} is Misspecified	20
6	Real Data	22
6.1	Assessing Goodness of Fit	23
7	Discussion	24
A	Appendix: Computing Joint Density for a MRF	28
B	More on Simulations	29

1 Introduction

Usually not including explicitly the sampling mechanism in the likelihood function does not have an impact in the inference (either likelihood-based or Bayesian) of a population quantity. All that is needed is the vector of indicators that encode which individuals have been included in the sample. Rubin [Rubin \(1976\)](#) and Heitjan and Rubin [Heitjan and Rubin \(1991\)](#) have developed a rigorous approach for tackling the question of when it is valid to ignore the functional form of the sampling mechanism for performing inferences. A key notion for this approach is the one of *ignorability*, which establishes when the probability distribution of the sampling design is relevant for modeling the distribution of random quantities corresponding to individuals not included in the sample. Under Rubin’s framework ignorability is equivalent to saying that the posterior of the population quantity can be computed without conditioning on the functional form of the sampling design. A sampling design is called *non-ignorable* if its functional form has to be expressed explicitly in the model in order to perform likelihood-based or Bayesian inference.

We consider a situation where non-ignorability arises because the sampling design is driven by a network, which is progressively discovered through sampling. One way to understand this is that the likelihood will depend on quantities indexed to the individuals not sampled. This could happen in at least two ways:

1. The probability distribution of the sampling design depends on features corresponding the portion of the graph that was not sampled. An obvious implication of this is that changes in the underlying network will affect the likelihood. An equally important, but greatly overlooked aspect of this is that we could have different realizations of the sampling mechanism, leading to the same set of sampled individuals, but conveying different information about the network structure, therefore producing different inferences.
2. The network induces a dependence structure on the responses, in such case the responses of not sampled individuals need to be taken into account for computing the likelihood.

The first point is illustrated in [Figure 1](#). It shows the case when the likelihood of the sample is affected by the underlying network and the case where the same set of sampled individuals can arise from different realizations of the design implying different values for the likelihood.

From a statistical perspective, the issues of inferring a population quantity using a non-ignorable sampling design on a social network include: Modeling the unobserved part of the social network in probabilistic fashion, understanding the sampling mechanism as a probability model and including it in the likelihood, and finally, modeling the dependence structure of the responses given the

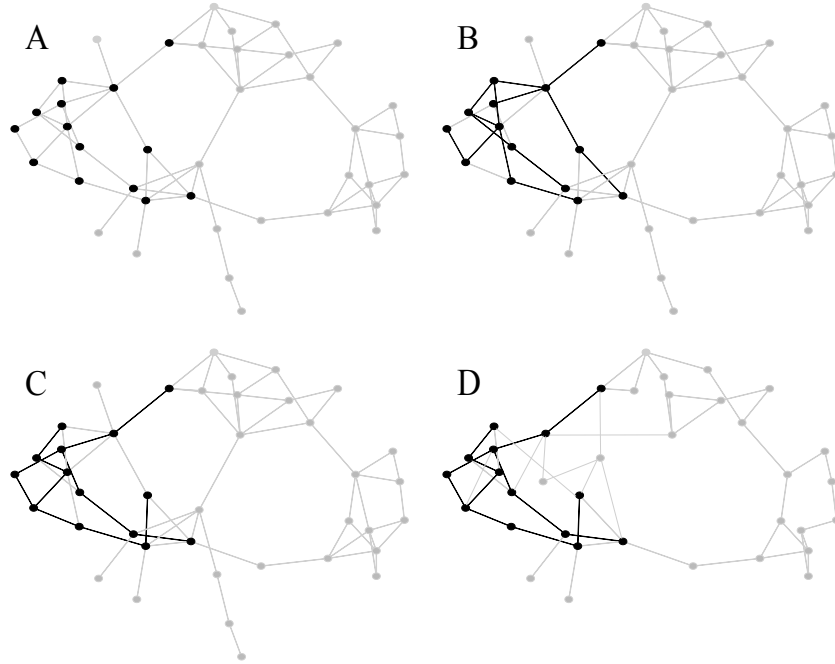


Figure 1: An illustration of the sources of non-ignorability in link-tracing designs. *Panel A:* A population graph and a random sample. *Panel B:* A realization of a link-tracing design on that produces the sample in panel A. *Panel C:* A different realization of a link-tracing design on that produces the same sample in panel A. *Panel D:* The same realization of a link-tracing design in panel C, but on a different population graph; it has a different likelihood.

network. Not taking into account any of these issues leads to misleading inferences and drastic understatements regarding the uncertainty associated to those inferences. This has been examined throughly in (Cite Michael O).

A sample design based on a social network structure that is widely used for inferring a population quantity is Respondent-Driven Sampling (RDS). This design was proposed by Heckathorn [Heckathorn \(1997\)](#). Later Volz and Heckathorn [Volz and Heckathorn \(2008\)](#) proposed an estimation procedure tailored for RDS based on the assumption that the relationship between the probability of inclusion of a given individual and the degree of the corresponding node is linear. Gile [Gile \(2011\)](#) improved this methodology by estimating the relationship between inclusion probability and degree distribution via an iterative procedure. None of this approaches is model-based, therefore, they are all vulnerable to the issues mentioned before. What is more relevant to the discussion is that these approaches assume that RDS is an ignorable design; we prove this is not true. A very interesting work that involves the concept of ignorability is Hancock and Gile [Hancock and Gile \(2011\)](#). Their focus is on estimating the parameters of the social network model, not in estimating a population quantity while allowing uncertainty for the network structure.

The purpose of this paper is to propose methodology such that:

1. Allows us to perform inferences in cases where the sampling mechanism is non-ignorable.
2. Takes into account all relevant sources of uncertainty, *i.e.*, uncertainty regarding the underlying social network, the sampling mechanism and the parameters of the model.
3. Models the dependence structure of the response explicitly by using the social network structure.
4. Is modular, in the sense that different priors and likelihoods can be used for the components of the model.

We developed a general framework that accomplishes these four objectives. The underlying network structure is understood as a random graph model. The dependence structure of the responses given the social network is modeled via a Markov Random Field (MRF). We write likelihoods for the sampling mechanisms given the graph. Inferences on the model are performed using a Bayesian approach [Robert \(2001\)](#). An important contribution of this paper is that it departs from the generality of Rubin’s framework in the sense that instead of considering a joint distribution for all the elements in the model with no structure at all, we impose a series of conditional independence statements among them that correspond to reasonable general assumptions about sampling on social networks problems.

A very interesting feature of our framework is that it allows us to circumvent the assumptions established in [Salganik and Heckathorn \(2004\)](#) and [Heckathorn \(2007\)](#) in order to guarantee consistency of the Volz-Heckathorn estimator for RDS sampling. We elaborate on this point in the discussion.

The outline of the paper goes as follows: In Section [2](#) we phrase the problem in terms of the general framework established by [Rubin \(1987\)](#) and set the basis for talking rigorously about the role of knowing the distribution of the sampling mechanism for performing inferences on the population quantity. In Section [3](#) we present a general framework for performing Bayesian inferences about a population quantity in a social network context. We also provide specific choices for the distributions required by the model and a description of the Markov Chain Monte Carlo (MCMC) scheme for this model. Simulations illustrating the advantages of the proposed methodology are shown in Section [5](#). Section [7](#) is the discussion.

2 Theory and Definitions

2.1 The Concept of Ignorability in the Context of Social Networks

We are interested in performing Bayesian inference on a population quantity Q . We assume that $Q = Q(X, W)$, where:

1. W denotes the *response vector*, i.e. those variables that are of primary interest to the investigator (in the sense that the scientific question at hand makes direct reference to them). To observe these variables, it is a necessary condition that the corresponding individual is included in the sample.
2. X denotes a vector of covariates that is available for all individuals in the population, regardless if they are included in the sample or not.

This setting was established in [Rubin \(1987\)](#). We denote the sampling mechanism by I . Here I plays two roles: it represents the sampling mechanism (the dynamic process of data collection) as a probability model and the indicator function for a specific sample. For the latter, we adopt the convention of setting $I = 1$ for the individuals that were included in the sample and $I = 0$ for those who were not included.

Ignorability [Rubin \(1987\)](#) is a property of the sampling mechanism I with respect to a model $p(W, X, I)$. A sampling mechanism is called *ignorable* with respect to a model $p(W, X, I)$ if:

$$\Pr(I \mid W, X) = \Pr(I \mid W_{INC}, X), \quad (1)$$

where W_{INC} is the subset of W for which $I = 1$. The subset of W for which $I = 0$ is denoted by W_{EXC} . I is called *non-ignorable* if Expression 1 does not hold. The statement made in Expression 1 is equivalent to

$$\Pr(W_{EXC} \mid W_{INC}, X, I) = \Pr(W_{EXC} \mid W_{INC}, X), \quad (2)$$

this is, that the uncertainty in W_{EXC} , and therefore, the uncertainty in Q can be expressed in terms of the observed data and the indicator that specifies which individuals are included in the sample only, without the need of making reference to the functional form of the sampling mechanism.

Two important concepts that are related to this notion of ignorability are the one of *ignorable missing data mechanism* and the one of *amenability*. Both of them are based in a factorization of the form:

$$p(W, X, I) = p(W, X \mid \theta)p(I \mid W, \eta), \quad (3)$$

and rely on the assumption that θ and η distinct. In both settings it is assumed that the data are *missing at random*, which is the property encoded in Equation 2. The difference between the notion of ignorability given by Expression 1 and these two concepts is that the former is used for the inference of a population quantity, while the later are used when the goal is to make inferences on θ , the parameter that drives the distribution of the full data.

For the purpose of this paper, it is reasonable to set $W = (Y, \mathcal{G})$, where \mathcal{G} is the graph or social network and Y is the vector of univariate responses associated to each node of \mathcal{G} . This means that a sampling mechanism I is non-ignorable if its distribution is not constant with respect to unobserved data (unobserved entries of Y or features of the unobserved portion of the network). We denote by \mathcal{G}_{INC} and \mathcal{G}_{EXC} , respectively, the observed and unobserved parts of the network. Y_{INC} denotes the responses of the individuals that were included in the sample, Y_{EXC} denotes the responses of the individuals that were not included in the sample.

2.2 Respondent-Driven Sampling

Respondent-Driven Sampling (RDS) is a sampling procedure proposed by Heckathorn [Heckathorn \(1997\)](#) that consists in picking r individuals according to some deterministic mechanism and propagating the sample using a policy of coupons that takes advantage of the social network \mathcal{G} . We describe the RDS algorithm briefly:

1. A set of r nodes (where $r < n$) are selected via a deterministic mechanism. These r nodes are known as *seeds*.
2. Each individual corresponding to a seed is given m coupons. Each of these individuals gives the coupons according to a uniform distribution over her (his) neighbors in \mathcal{G} that have not been sampled already. The nodes recruited this way are known as the *first wave*.
3. Each node sampled at the $k - 1$ wave ($k \geq 2$) is given m coupons. Each of the individuals corresponding to this nodes gives the coupons according to a uniform distribution over her (his) neighbors in \mathcal{G} that have not been sampled already. The nodes recruited this way are known as the *k-th wave*.
4. If the number of available neighbors (*i.e.*, not sampled already) of any sampled node, is less or equal to m , then all of them receive coupons. This applies to every wave.
5. Continue sampling nodes until a pre-specified sample size n is attained. Waves do not have to be completed.

Let I denote RDS, which is understood as a probabilistic process that propagates through a network \mathcal{G} . The probability distribution for I given \mathcal{G} can be written as:

$$p(I \mid \mathcal{G}) = \frac{1}{\binom{\tilde{d}_0}{m}} \left(\prod_{j_1=1}^{w_0} \frac{1}{\binom{\tilde{d}_{j_1}}{m}} \left[\prod_{j_2, j_1=1}^{w_{j_1}} \frac{1}{\binom{\tilde{d}_{j_2, j_1}}{m}} \cdots \left[\prod_{j_k, \dots, j_1=1}^{w_{j_{k-1}, \dots, j_1}} \frac{1}{\binom{\tilde{d}_{j_k, \dots, j_1}}{m}} \right] \cdots \right] \right).$$

Here $\tilde{d}_{(\cdot)}$ stands for the number of neighbors (*i.e.*, adjacent nodes with respect to \mathcal{G}) of the corresponding node that have not been sampled yet; $w_{(\cdot)}$ denotes the number of recruited neighbors by a given node during the previous wave, while k denotes the number of waves needed to recruit n individuals (such number is a function of the RDS policies and the restrictions imposed by the graph topology).

Respondent-Driven Sampling is not ignorable. The distribution of the sampling mechanism I depends on the subgraph of \mathcal{G} that is included in the sample, but it also depends on information regarding \mathcal{G}_{EXC} . If a node is recruited and has d neighbors, where $d > m$, m being the number of coupons per individual, then the distribution of I depends on the number of neighbors of the node not included in the sample. Then,

$$\Pr(I \mid Y, \mathcal{G}) \neq \Pr(I \mid Y_{INC}, \mathcal{G}_{INC}).$$

It follows that I is not ignorable.

3 Statistical Methodology

3.1 Modeling framework

We assume a data generative model of the form:

$$p(Y, I, \mathcal{G}, \alpha, \gamma) = p(\alpha)p(\mathcal{G} \mid \alpha)p(I \mid \mathcal{G})p(\gamma)p(Y \mid \mathcal{G}, \gamma). \quad (4)$$

This means that we understand \mathcal{G} as a realization of a random graph model with parameter vector α . As explained before, the distribution of I is conditional on a given realization of \mathcal{G} , it is not necessary to know how the graph was generated to sample I (an assumption of link-tracing designs). Therefore α does not appear on the conditional for I . A key assumption of our approach is that \mathcal{G} induces a dependence structure on Y . To fully specify such dependence structure, additional parameters may be needed; those parameters are denoted by γ , that explains the term $p(Y \mid \mathcal{G}, \gamma)$.

$p(\alpha)$ and $p(\gamma)$ are the priors for α and γ , respectively.

Observe that we include the factor $p(I \mid \mathcal{G})$ to deal with non-ignorability issues. Still, to achieve this, we need to model the underlying graph: that is the purpose of adding the factor $p(\mathcal{G} \mid \alpha)p(\alpha)$. For this paper we assume that the dependence structure modeled by $p(Y \mid \mathcal{G}, \gamma)$ is a Markov Random Field (MRF).

Note that Expression 4 is compatible with the factorization described in Expression 3. Here $W = (Y, \mathcal{G})$ and $\theta = (\alpha, \gamma)$. For the problems discussed in this paper, the sampling mechanism will be driven by tuning parameters that are known to the researcher, and therefore, they are not part of the inference. For this reason there is no equivalent to η in Expression 4. Also, for the sake of simplicity, we do not include covariates (X in Expression 3), but there is no impediment for incorporating them if needed.

3.2 Posterior inference and estimation

Because of the MRF assumption, to compute the likelihood of Y_{INC} it is necessary to augment versions of $(Y_{EXC}, \mathcal{G}_{EXC})$ to the observed data. Augmenting \mathcal{G}_{EXC} is also a key step for dealing with the non-ignorability of the sampling mechanism (as explained in Sections 2.2 and 3.1). We denote by Y_{AUG} and \mathcal{G}_{AUG} those Monte Carlo versions. Since we do not know how many nodes and edges were unobserved due to I , the model becomes one of variable dimension. Let us introduce further notation: We denote by N_{AUG} the number of augmented nodes, N_{MC} denotes the sum of N_{AUG} and n (the sample size). Finally, we denote by \mathcal{G}_{MC} the union of \mathcal{G}_{INC} and \mathcal{G}_{AUG} . For the sampling mechanisms we considered for this paper, the way \mathcal{G}_{INC} is augmented will have an impact on the factor of the likelihood that includes I , *i.e.*, they are non-ignorable.

To perform inference on the model we just proposed we will pursue a Bayesian model averaging (see Raftery et al. (1996) and Robert (2001), Section 7.4) strategy:

$$p(Q \mid Y_{INC}, \mathcal{G}_{INC}, I) = \sum_w p(\mathcal{M}_w) \int_{\Theta_w} p_w(Q \mid \theta_w) p(\theta_w \mid Y_{INC}, \mathcal{G}_{INC}, I) d\theta_w.$$

Where $\theta_w = (Y_{AUG}(w), \mathcal{G}_{AUG}(w), \alpha(w), \gamma(w))$ and \mathcal{M}_w is the set of graphs with given number of augmented nodes and edges. The rationale behind this approach is that the information regarding how to augment \mathcal{G}_{INC} is not contained in the data $(Y_{INC}, \mathcal{G}_{INC}, I)$, but it is encoded in the prior $p(\alpha)p(\mathcal{G} \mid \alpha)$. Once the number of augmented number of nodes and edges has been specified, samples from the conditional model $p(\theta_w \mid Y_{INC}, \mathcal{G}_{INC}, I)$ can be obtained using standard Markov Chain Monte Carlo (MCMC) techniques.

The way $p(\mathcal{M}_w)$ is constructed goes as follows:

1. We obtain D samples from $p(\mathcal{G}, \alpha) = p(\alpha)p(\mathcal{G} | \alpha)$.
2. For each sample, a realization of $(I | \mathcal{G})$ is obtained.
3. This implies a Monte Carlo version of $(\mathcal{G}_{INC}, \mathcal{G}_{EXC})$. From here, the number of nodes and edges to be augmented can be computed.

Since this is done D times, the distribution of the number of augmented nodes and edges can be calibrated.

For the augmented edges we consider two cases:

1. Edges that are incident only to nodes for which $I = 1$. For those edges we impose the restriction that the corresponding nodes must belong to different waves of the link-tracing design. The rationale of this condition will be clear once we describe the part of the MCMC devoted to update \mathcal{G} .
2. Edges that are incident to only one sampled node.

We are discarding the case where edges are incident only to augmented nodes. Note that such edges would not contribute to the likelihood of the sampling mechanism $p(I | \mathcal{G})$ at all, in contrast, adding them would make the update of \mathcal{G} much more difficult. For each augmented node, we require it to be connected to a sample node via an augmented edge. Otherwise there would not be a contribution to $p(I | \mathcal{G})$. We acknowledge that such restrictions imply a loss of information when performing inference on $p(Y | \mathcal{G}, \gamma)$; We elaborate on this point in the discussion. When sampling from $p(\mathcal{M}_w)$ three numbers are obtained: the number of nodes to augment, the number of edges to augment such that they are incident to sampled nodes only (we call these *intra-sample edges*), and the edges incident to one sampled node only (let us call them *extra-sample edges*).

An interesting feature of this problem is that the parameter of interest Q is not present explicitly in the model. Clearly the probability of the response taking the value 1 is not uniform across individuals, therefore Q should be understood as the mean probability of the response taking the value 1. With this in mind we propose the following estimator: For each iteration of the MCMC we obtain the vector:

$$Y_{MC} = (Y_{INC}, Y_{AUG}).$$

Let Q_{MC} be the sample mean of Y_{MC} . The point estimator we propose \hat{Q}_B is the sample mean of the Q_{MC} 's. The credible intervals will be given by the corresponding empirical quantiles of the

Q_{MC} 's. Note that the Q_{MC} 's have the same meaning for the models originated from the different graph complexities generated by $p(\mathcal{M}_w)$, therefore it makes sense to combine posterior samples using BMA.

3.3 Model Specification

We now present the specific choices we made for these distributions: For \mathcal{G} an Erdos-Renyi model [Erdos and Renyi \(1960\)](#) was assumed, with a single probability of inclusion $\alpha \in (0, 1)$. A Beta(ω_1, ω_2) was used as prior for α . Our specification for $p(I \mid \mathcal{G})$ is an RDS with m coupons per wave and sample size n . For this paper we will assume that y_i is binary, and

$$\Pr\{y_i = 1 \mid Y_{-i}, \mathcal{G}, \gamma\} = \Phi \left(\psi + \sum_{\{k \mid A(i,k)=1\}} \zeta y_k \right), \quad (5)$$

where A is the adjacency matrix for \mathcal{G} and $\gamma = (\psi, \zeta)$. In other words $p(Y \mid \mathcal{G}, \gamma)$ is specified as a Markov Random Field (MRF) based on a probit model. We used a scaled Beta as prior for ζ , *i.e.*

$$p(\zeta \mid \eta_1, \eta_2, \delta) = \frac{1}{B(\eta_1, \eta_2)} \frac{\zeta^{\eta_1-1} (\delta - \zeta)^{\eta_2-1}}{\delta^{\eta_1+\eta_2-1}} \times \mathbb{I}_{(0,\delta)}(\zeta), \quad (6)$$

and a density of the form:

$$p(\psi \mid \nu_1, \nu_2, \xi) = \frac{1}{B(\nu_1, \nu_2)} \frac{(\psi + \xi)^{\nu_1-1} (-\psi)^{\nu_2-1}}{\xi^{\nu_1+\nu_2-1}} \times \mathbb{I}_{(-\xi,0)}(\psi) \quad (7)$$

as prior for ψ .

We adopt these distributions for the sake of concreteness; they are not essential to our methodology. Any sampling design that propagates through a social network could be used to specify $p(I \mid \mathcal{G})$. In principle, we could use any random graph model for $p(\mathcal{G} \mid \alpha)$, as long as:

1. The density of the random graph model can be computed efficiently.
2. It is feasible to marginalize efficiently, this is, to obtain the distribution of \mathcal{G}_{INC} given any realization of I .

For instance, we also consider the following prior:

$$\varphi_1, \dots, \varphi_N \sim B(\alpha_1, \alpha_2) \quad (8)$$

$$A(i, k) \sim \text{Ber}(\varphi_i \varphi_k), \quad (9)$$

which is inspired by the model proposed by [Perry and Wolfe](#). The choice of a specific MRF is more delicate, since it involves drastic changes in the MCMC procedure. Still, all these considerations are computational, not conceptual.

4 Markov Chain Monte Carlo Algorithm

First, let us define some notation:

1. We denote by n the sample size;
2. We denote by N_{AUG} the number of augmented nodes;
3. Let $N_{MC} = N_{AUG} + n$;
4. Let \mathcal{G}_{MC} be the graph obtained from merging \mathcal{G}_{INC} and \mathcal{G}_{AUG} .
5. We use the superscript (t) to denote the t -th iteration of the MCMC.

We use a mixture of kernels to perform the updates on the different components of the model. Regarding the updates for \mathcal{G}_{AUG} , we impose the following policies:

1. For the edges that are incident only to nodes for which $I = 1$. For those edges we impose the restriction that the corresponding nodes must belong to different waves of the link-tracing design. The rationale of this condition will be clear once we describe the part of the MCMC devoted to update \mathcal{G} .
2. For the edges that are incident to only one sampled node, we impose no restrictions.
3. We are discarding the case where edges are incident only to augmented nodes. Note that such edges would not contribute to the likelihood of the sampling mechanism $p(I \mid \mathcal{G})$ at all, in contrast, adding them would make the update of \mathcal{G} much more difficult. For each augmented node, we require it to be connected to a sample node via an augmented edge. Otherwise there would not be a contribution to neither $p(Y \mid \mathcal{G}, \gamma)$ nor $p(I \mid \mathcal{G})$.

We now describe the elements of the mixture in the following subsections. In the last subsection we describe the proposal distributions used in this paper.

4.1 Update ψ

Let $q(\cdot | \cdot)$ denote the proposal distribution; it follows that the Metropolis ratio is of the form:

$$\frac{p(\psi^{(t+1)})p(Y | \mathcal{G}, \zeta, \psi^{(t+1)})}{p(\psi^{(t)})p(Y | \mathcal{G}, \zeta, \psi^{(t)})} \times \frac{q(\psi^{(t)} | \psi^{(t+1)})}{q(\psi^{(t+1)} | \psi^{(t)})}.$$

The part of the quotient regarding the proposal distribution and the prior for ψ is easy to handle. We focus our discussion on the part concerning de joint $p(Y | \mathcal{G}, \zeta, \psi)$, which is a MRF based on a probit (Expression 5). We rewrite that part of the quotient in the following way:

$$\frac{p(Y | \mathcal{G}, \zeta, \psi^{(t+1)})}{p(Y | \mathcal{G}, \zeta, \psi^{(t)})} = \frac{p(Y | \mathcal{G}, \zeta, \psi^{(t+1)}) \div p(0 | \mathcal{G}, \zeta, \psi^{(t+1)})}{p(Y | \mathcal{G}, \zeta, \psi^{(t)}) \div p(0 | \mathcal{G}, \zeta, \psi^{(t)})} \times \Lambda. \quad (10)$$

We proceed this way because the quotient on the left side of Equation 10 cannot be computed directly. The idea is that the two quotients on the right side of Equation 10 and the term denoted by Λ can be computed in an efficient manner. For the quotients this is easy to prove by using an argument very similar to Brook's Lemma Brook (1964):

$$\frac{p(Y | \mathcal{G}, \zeta, \psi)}{p(0 | \mathcal{G}, \zeta, \psi)} = \frac{p(y_1, y_2, \dots, y_{N_{MC}} | \mathcal{G}, \zeta, \psi)}{p(0, y_2, \dots, y_{N_{MC}} | \mathcal{G}, \zeta, \psi)} \times \frac{p(0, y_2, \dots, y_{N_{MC}} | \mathcal{G}, \zeta, \psi)}{p(0, 0, \dots, y_{N_{MC}} | \mathcal{G}, \zeta, \psi)} \dots \times \frac{p(0, 0, \dots, 0, y_{N_{MC}} | \mathcal{G}, \zeta, \psi)}{p(0, 0, \dots, 0 | \mathcal{G}, \zeta, \psi)}.$$

Each of the terms on the right side can be computed efficiently by applying the identity:

$$\frac{p(y_1, y_2, \dots, y_k, \dots, y_{N_{MC}} | \mathcal{G}, \zeta, \psi)}{p(y_1, y_2, \dots, y_k^*, \dots, y_{N_{MC}} | \mathcal{G}, \zeta, \psi)} = \frac{p(y_k | Y_{-k}, \mathcal{G}, \zeta, \psi)}{p(y_k^* | Y_{-k}, \mathcal{G}, \zeta, \psi)},$$

see Kaiser and Cressie (2000). The way to compute the Metropolis ratio will become explicit once we determine Λ . We start by expanding the terms in the right side of Expression 10:

$$\frac{p(Y | \mathcal{G}, \zeta, \psi^{(t+1)})}{p(Y | \mathcal{G}, \zeta, \psi^{(t)})} \times \left(\frac{1 - \Phi(\psi^{(t)})}{1 - \Phi(\psi^{(t+1)})} \right)^{N_{AUG}} \frac{p(0 | \mathcal{G}_{INC}, \zeta, \psi^{(t)})}{p(0 | \mathcal{G}_{INC}, \zeta, \psi^{(t+1)})} \times \Lambda \quad (11)$$

This follows from assuming a probit model for the full conditionals of the MRF and the restriction of \mathcal{G}_{MC} which forbids the existence of edges incident to augmented nodes only. By a similar argument and now using the restriction on \mathcal{G}_{MC} that there will be no augmented edges connecting nodes from different waves, we obtain that Expression 11 is equal to:

$$\frac{p(Y | \mathcal{G}, \zeta, \psi^{(t+1)})}{p(Y | \mathcal{G}, \zeta, \psi^{(t)})} \times \left(\frac{1 - \Phi(\psi^{(t)})}{1 - \Phi(\psi^{(t+1)})} \right)^{N_{MC}-1} \frac{p(y_1 = 0 | \zeta, \psi^{(t)})}{p(y_1 = 0 | \zeta, \psi^{(t+1)})} \times \Lambda. \quad (12)$$

Therefore:

$$\Lambda = \frac{p(y_1 = 0 \mid \zeta, \psi^{(t+1)})}{p(y_1 = 0 \mid \zeta, \psi^{(t)})} \times \left(\frac{1 - \Phi(\psi^{(t+1)})}{1 - \Phi(\psi^{(t)})} \right)^{N_{MC}-1}.$$

The terms of the form $p(y_1 = 0 \mid \zeta, \psi)$ can be computed analytically: We consider the subgraph of \mathcal{G}_{MC} induced by the star of the node associated to y_1 , then the MRF corresponding to that subgraph, ζ and ψ is defined. That MRF is small enough so that the methodology proposed by Kaiser and Cressie [Kaiser and Cressie \(2000\)](#) can be applied and the joint distribution of the y_i 's associated to the star can be obtained. After that computing the marginal probability of $\{y_i = 0\}$ is trivial.

4.2 Update ζ

Let us denote the proposal distribution by $q(\cdot \mid \cdot)$. The Metropolis ratio is given by:

$$\frac{p(\zeta^{(t+1)})p(Y \mid \mathcal{G}, \zeta^{(t+1)}, \psi)}{p(\zeta^{(t)})p(Y \mid \mathcal{G}, \zeta^{(t)}, \psi)} \times \frac{q(\zeta^{(t)} \mid \zeta^{(t+1)})}{q(\zeta^{(t+1)} \mid \zeta^{(t)})}.$$

As we discussed for the update for ψ , our focus will be on the part of the ratio that involves the MRF density $p(Y \mid \mathcal{G}, \zeta, \psi)$. We proceed the same way as in the case for ψ ; the only change here is that the terms of the form $(1 - \Phi(\psi))$ cancel out in Expression 12. Therefore:

$$\Lambda = \frac{p(y_1 = 0 \mid \zeta^{(t+1)}, \psi)}{p(y_1 = 0 \mid \zeta^{(t)}, \psi)}.$$

4.3 Update Y_{AUG}

We denote the proposal distribution by $q(\cdot \mid \cdot)$. The Metropolis ratio is given by:

$$\frac{p(Y_{INC}, Y_{AUG}^{(t+1)} \mid \mathcal{G}, \zeta, \psi)}{p(Y_{INC}, Y_{AUG}^{(t)} \mid \mathcal{G}, \zeta, \psi)} \times \frac{q(Y_{AUG}^{(t)} \mid Y_{AUG}^{(t+1)})}{q(Y_{AUG}^{(t+1)} \mid Y_{AUG}^{(t)})}.$$

Note that the joint density of the MRF can be expressed as:

$$p(Y_{AUG} \mid Y_{INC}, \mathcal{G}, \zeta, \psi) \times p(Y_{INC} \mid \mathcal{G}, \zeta, \psi)$$

It follows that the part of the Metropolis ratio involving the quotient of MRF densities can be simplified as:

$$\frac{p(Y_{AUG}^{(t+1)} \mid Y_{INC}, \mathcal{G}, \zeta, \psi)}{p(Y_{AUG}^{(t)} \mid Y_{INC}, \mathcal{G}, \zeta, \psi)} \tag{13}$$

Now we will use the restriction which forbids the existence of edges incident to augmented nodes only. Expression 13 is equivalent to:

$$\frac{\prod_{k=1}^{N_{AUG}} p(Y_{AUG(k)}^{(t+1)} | Y_{INC}, \mathcal{G}, \zeta, \psi)}{\prod_{k=1}^{N_{AUG}} p(Y_{AUG(k)}^{(t)} | Y_{INC}, \mathcal{G}, \zeta, \psi)}.$$

The product is over the conditionals for the augmented nodes only ($AUG(k)$ is the index for the k -th augmented node). This implies that the part of the Metropolis ratio involving MRF densities can be written in terms of the full conditionals, which can be computed efficiently. For concreteness we specified the full conditionals using the probit model (Expression 5).

4.4 Update \mathcal{G}_{AUG}

Let q denote the proposal distribution; the Metropolis ratio for updating \mathcal{G}_{AUG} is of the form:

$$\frac{p(\mathcal{G}^{(t+1)} | \alpha) p(I | \mathcal{G}^{(t+1)}) p(Y | \mathcal{G}^{(t+1)}, \zeta, \psi)}{p(\mathcal{G}^{(t)} | \alpha) p(I | \mathcal{G}^{(t)}) p(Y | \mathcal{G}^{(t)}, \zeta, \psi)} \times \frac{q(\mathcal{G}^{(t)} | \mathcal{G}^{(t+1)})}{q(\mathcal{G}^{(t+1)} | \mathcal{G}^{(t)})}.$$

As with Y_{AUG} , ζ and ψ , we focus on the factor of the Metropolis ratio involving MRF densities. The way such factor is simplified depends on which part of \mathcal{G}_{AUG} will be updated. To update the extra-sample edges we follow a very similar reasoning to the one used for Y_{AUG} , this is, the factor

$$\frac{p(Y | \mathcal{G}^{(t+1)}, \zeta, \psi)}{p(Y | \mathcal{G}^{(t)}, \zeta, \psi)} \quad (14)$$

can be written as

$$\frac{\prod_{k=1}^{N_{AUG}} p(Y_{AUG(k)} | Y_{INC}, \mathcal{G}^{(t+1)}, \zeta, \psi)}{\prod_{k=1}^{N_{AUG}} p(Y_{AUG(k)} | Y_{INC}, \mathcal{G}^{(t)}, \zeta, \psi)}.$$

To compute the Metropolis ratio the update of intra-sample edges, we expand Expression 14 in the following way:

$$\frac{p(Y_{AUG} | Y_{INC}, \mathcal{G}^{(t+1)}, \zeta, \psi)}{p(Y_{AUG} | Y_{INC}, \mathcal{G}^{(t)}, \zeta, \psi)} \times \frac{p(Y_{INC} | \mathcal{G}^{(t+1)}, \zeta, \psi)}{p(Y_{INC} | \mathcal{G}^{(t)}, \zeta, \psi)}. \quad (15)$$

We now introduce some notation: Let Y_{INC^*} be the values of the response corresponding to the sampled nodes ($I = 1$) that are adjacent to the augmented nodes according to \mathcal{G} . Denote by \mathcal{G}_{INC^*} the induced graph of \mathcal{G} obtained by taking the augmented nodes and the ones adjacent to them. Clearly the first factor in Expression 15 is equal to:

$$\frac{p(Y_{AUG} | Y_{INC^*}, \mathcal{G}_{INC^*}^{(t+1)}, \zeta, \psi)}{p(Y_{AUG} | Y_{INC^*}, \mathcal{G}_{INC^*}^{(t)}, \zeta, \psi)}. \quad (16)$$

Because we are updating the intra-sample edges, Expression 16 is equal to 1. We introduce further notation: Let $Y_{INC}[i]$ denote the values of the response for the sampled nodes recruited in waves previous to the i th (including the seed). We denote by $\mathcal{G}[i]$ the subgraph of \mathcal{G} induced by the nodes recruited in waves 1 to i (including the seed). The terms in the second factor of Expression 15 can be expanded as:

$$\left(\prod_{i=1}^{N_W} \prod_{k=1}^{N_W(i)} p(Y_{INC}(i, k) \mid Y_{INC}[i], \mathcal{G}[i], \zeta, \psi) \right) \times p(Y_{INC}[0] \mid \zeta, \psi);$$

Here $Y_{INC}[0]$ denotes the seed, N_W is the total number of waves, and $N_W(i)$ is the number of nodes recruited in the i -th wave. This implies that Expression 14 can be simplified as follows:

$$\frac{\prod_{i=1}^{N_W} \prod_{k=1}^{N_W(i)} p(Y_{INC}(i, k) \mid Y_{INC}[i], \mathcal{G}^{(t+1)}[i], \zeta, \psi)}{\prod_{i=1}^{N_W} \prod_{k=1}^{N_W(i)} p(Y_{INC}(i, k) \mid Y_{INC}[i], \mathcal{G}^{(t+1)}[i], \zeta, \psi)},$$

where some terms will cancel out.

4.5 Proposals

The proposals we used to update each parameter ($Y_{AUG}, \mathcal{G}_{AUG}, \psi, \zeta$) are the following:

1. For ζ we use a mixture kernel, where one component is a random walk reflecting at 0 and δ , the second component is the prior (Equation 6) and the third is a uniform distribution over $(0, \delta)$.
2. For ψ we implement a mixture kernel, where one component is a random walk reflecting at $-\xi$ and 0, the second component is the prior (Equation 7) and the third is a uniform distribution over $(-\xi, 0)$.
3. For the extra-sample edges in \mathcal{G}_{AUG} we take each augmented node k and count its number of neighbors h_k according to the current version of \mathcal{G} , then we delete the corresponding edges and then obtain a random sample of size h_k over the nodes for which $I = 1$. New edges are added which connect those h_k nodes to k .
4. For the intra-sample nodes in \mathcal{G}_{AUG} we first delete all the current intra-sample edges and then add a new set of edges at random while respecting the restriction that no edge will exist between nodes included in the same wave and preserving the current density of the graph.

5. For Y_{AUG} we use a mixture kernel. In one of them we impute using independent Bernoullis with mean equal to the average of Y_{INC} . For the other we sample Y_{AUG} from the conditional distribution $p(Y_{AUG} \mid Y_{INC}, \mathcal{G}, \zeta, \psi)$.

5 Results

5.1 Simulation Study

We conducted a simulation study to gain better understanding of the performance of our method. First we considered regimes where the prior $p(\mathcal{G} \mid \alpha)p(\alpha)$ matches the mechanism that generated the data. To evaluate performance of point estimators, a Monte Carlo estimate of the bias was computed; to evaluate confidence intervals and credible regions, the frequency of coverage was used. The regimes for the simulation were determined by the following factors:

1. The density of the underlying network. We considered the values 0.1, 0.05, 0.1 and 0.2.
2. The number of referrals m used for the RDS sampling; We set $m \in \{3, 5\}$.
3. The size of the underlying network. We considered the values 100, 200 and 500.

The sample size was set as 50. An Erdős-Rényi model was used to generate the random graph data for all regimes, also, the parameters of the MRF were set up so $Q_\infty = 0.2$ for all scenarios. Here Q_∞ is understood as the mean of Y over repeated samples, given (α, ζ, γ) . We denote the sample mean of Y by Q_{Emp} . Our method was implemented using the distributions described in Section 3.3; this implies that an Erdős-Rényi prior for the underlying network was assumed when fitting the model. We compared our methodology to the Volz-Heckathorn estimator and the corresponding Bootstrap confidence interval. Results are summarized in Tables 1, 6 and 7.

For a given regime (namely $N = 200$, $n = 50$, $\alpha = 0.1$), we plotted the 95 per cent credible intervals associated to each simulation and the corresponding 95 per cent confidence intervals implied by bootstrapping Volz-Heckathorn (Figure 2). While both procedures generate estimates with similar bias and coverage, our method produces intervals that are at least half shorter in average. We also plotted the coverage against the average length of the (confidence or credible) interval for each method considering a the same set of regimes used in Tables 1, 6 and 7. Results are displayed in Figure 3

We also plotted the mean square error of both methods (Bayesian and Volz-Heckathorn); We did this first by considering it as a function of number of coupons in RDS (Figure 4), then as a

Density	n	m	Bias Q_∞	Coverage Q_∞	Bias Q_{Emp}	Coverage Q_{Emp}	Length	Method
0.01	50	3	-0.0062	0.93	-0.0052	0.93	0.12	Bayes
0.01	50	3	0.0017	0.89	0.0027	0.92	0.24	VH
0.05	50	3	-0.0043	0.92	-0.0001	0.93	0.11	Bayes
0.05	50	3	-0.0095	0.91	-0.0053	0.95	0.26	VH
0.1	50	3	0.0035	0.93	0.0073	0.91	0.11	Bayes
0.1	50	3	0.0018	0.90	0.0056	0.89	0.25	VH
0.2	50	3	0.0006	0.98	-0.0017	0.89	0.10	Bayes
0.2	50	3	0.0077	0.92	0.0090	0.92	0.23	VH

Table 1: Average bias, $Q - \hat{Q}$, and frequency of coverage for the Bayesian and non-model based approach. For the Bayesian method, the point estimator \hat{Q}_B is given by the posterior mean and the 95% credible region was used for interval estimation. We compared these summaries to the Volz-Heckartorn (VH) estimator and the 95% bootstrap confidence interval. The simulation scenarios are given by: Density of the underlying network, and the maximum number of referrals m . The size of the underlying network was set as 200 and the sample size as 50. 100 simulations were performed for each scenario. In all cases an Erdős-Rényi model is used to generate the data. For each simulation, the BMA was implemented using 5 samples from the mixing distribution; for each of these samples, an MCMC was run using 3, 000 for burn-in and 500 posterior samples.

function of the density of the underlying network (Figure 5), and finally as a function of sample size 6, for a fixed size of the underlying network (Figure). We observed that the mean square error is smaller for our method compared to VH for most regimes.

5.2 MCMC Performance

Let $h = h(\alpha, \gamma)$ be a generic function of the parameters that specify the random graph and the dependence structure. To compute an estimate of

$$\int \int h(\alpha, \gamma) p(\alpha, \gamma) d\gamma d\alpha,$$

an average of the posterior samples is used, *i.e.*,

$$S_T = \frac{1}{T} \sum_{t=1}^T h(\alpha^{(t)}, \gamma^{(t)}).$$

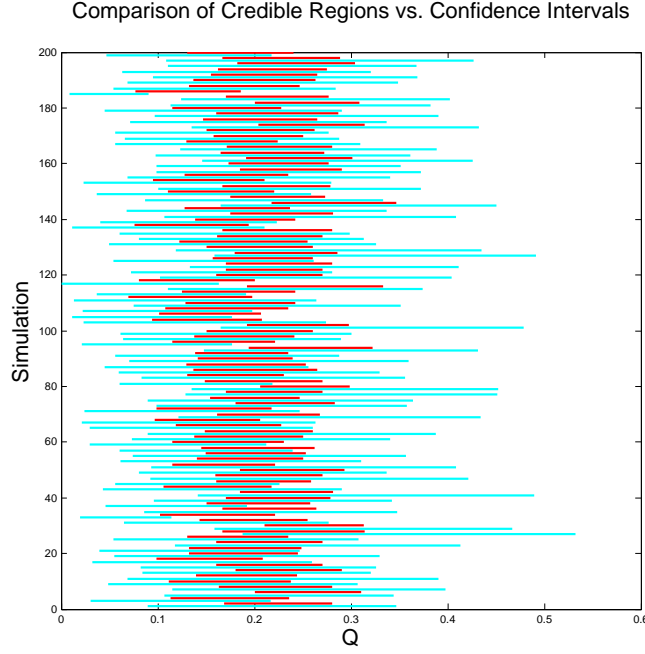


Figure 2: 95% credible intervals (red) vs. 95% bootstrap confidence intervals (cyan). For all data sets, the credible interval is plotted on top of the corresponding confidence interval. 100 Monte Carlo data sets were used to obtain this plot. We observed that our method produces intervals that are at least half shorter in average while having slightly higher coverage.

Here, T denotes the number of MCMC iterations after the burn-in period. The standard approach in the literature for estimating the variance of S_T is to use

$$\hat{\nu}_T = \frac{1}{T \times \hat{T}^S} \sum_{t=1}^T (h(\alpha^{(t)}, \gamma^{(t)}) - S_T)^2,$$

where \hat{T}^S denotes the *effective sample size*; this quantity is defined as:

$$\hat{T}^S = T/\kappa(h),$$

where

$$\kappa = 1 + 2 \sum_{t=1}^{\infty} \text{corr}(h(\alpha^{(0)}, \gamma^{(0)}), h(\alpha^{(t)}, \gamma^{(t)})).$$

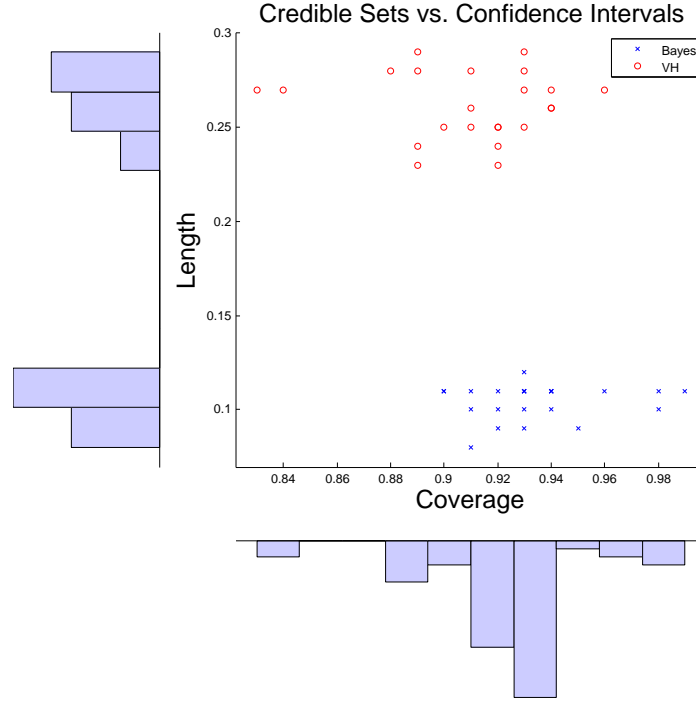


Figure 3: Average length vs. coverage of credible sets and confidence intervals for all regimes considered in Tables 1, 6 and 7. We observed that our method produces intervals that are at least half shorter in average while having similar coverage.

See [Liu and Chen \(1995\)](#) and Section 12.3.5 of [Robert and Casella \(2004\)](#).

To estimate the effective sample size of the MCMC procedure described in Section 4, we ran 200 independent chains per regime. The regimes were defined in terms of the size and density of the underlying network. T set as 1,000. The estimates for the effective sample size and the variance of S_T for different choices of h (to be precise: $h_1(\alpha, \gamma) = \psi$, $h_2(\alpha, \gamma) = \zeta$ and $h_3(\alpha, \gamma) = Q$) are displayed in Table 2. These results indicate that, for a network of size 100 and density between 0.01 and 0.2, our MCMC would generate roughly 20 independent samples for every 1,000 iterations.

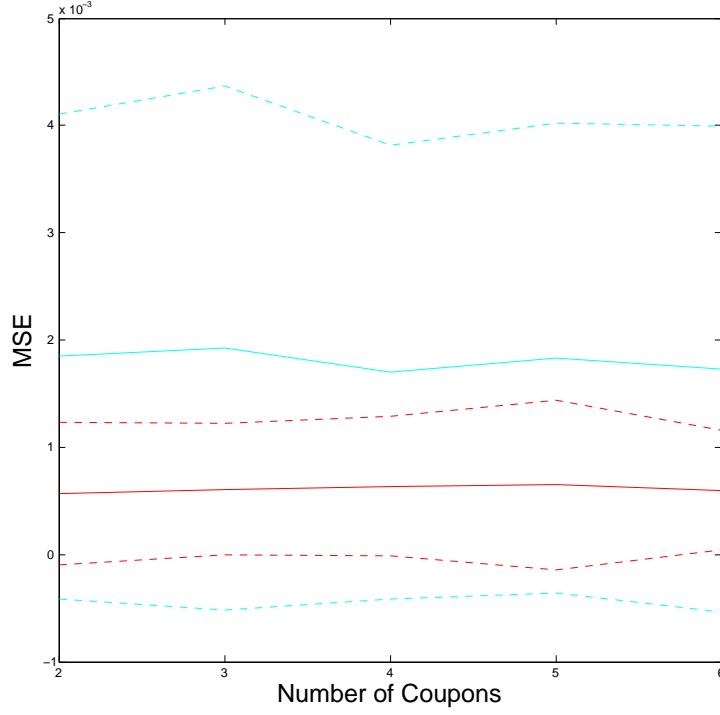


Figure 4: Mean square error as a function of number of coupons. The mean square error of our method as a function of coupons, plus and minus the standard deviation is displayed in red; for VH the curves are displayed in cyan. Here $N = 200$, $n = 50$ and $\alpha = 0.05$.

Density	N	$T/\kappa(h_1)$	$T/\kappa(h_2)$	$T/\kappa(h_3)$
0.01	100	11.18	15.78	21.33
0.05	100	10.56	14.73	20.89
0.1	100	10.27	14.46	20.53
0.2	100	10.12	14.32	20.14

Table 2: Effective sample size for $h_1(\alpha, \gamma) = \psi$, $h_2(\alpha, \gamma) = \zeta$ and $h_3(\alpha, \gamma) = Q$. Here $T = 1,000$. These results indicate that out of 1,000 MCMC iterations, 20 can be taken as roughly *iid.* of the posterior for Q , *i.e.*, it is recommended to apply lag 50 to the posterior samples.

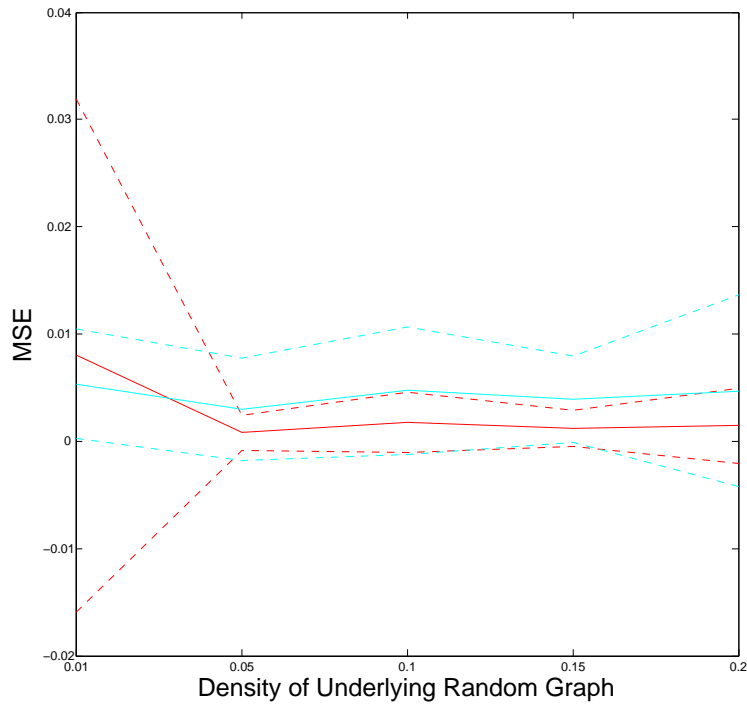


Figure 5: Mean square error as a function of density of the underlying network. The mean square error of our method as a function of density, plus and minus the standard deviation is displayed in red; for VH the curves are displayed in cyan. Here $N = 200$, $n = 50$. The number of coupons for RDS was set as 3.

5.3 When the Prior on \mathcal{G} is Misspecified

We ran one more simulation study where we considered regimes for which the prior $p(\mathcal{G} | \alpha)p(\alpha)$ had different functional form from the mechanism that generated the data. We evaluated the procedures using the same criteria as in the previous experiment. The regimes for the simulation were determined by the following factors:

1. The random graph model used to generate the underlying network was Small World.
2. The random graph models used to fit the model were Erdős-Rényi and Block-Model.

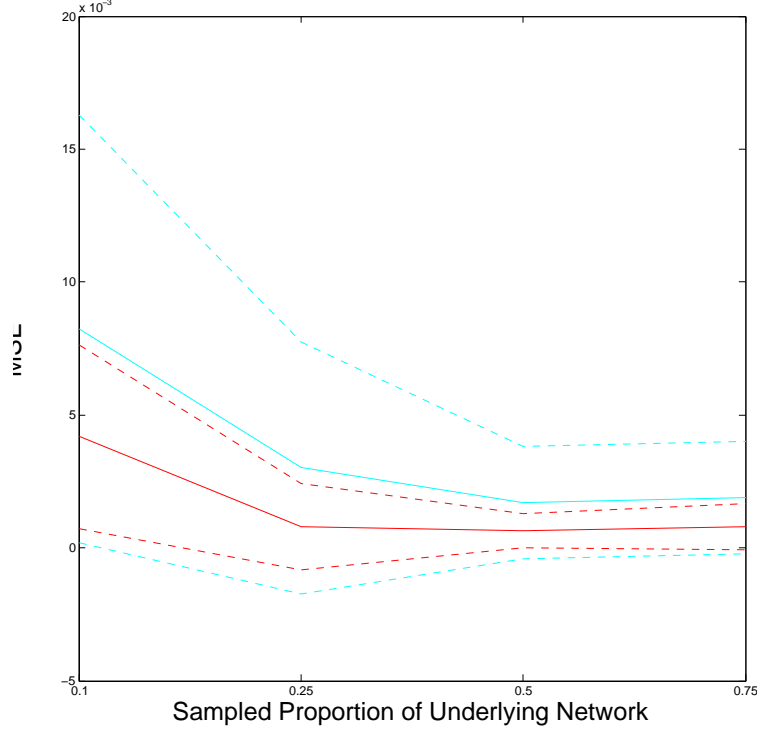


Figure 6: Mean square error as a function of the proportion of the underlying network included in the sample. The mean square error of our method as a function of sample size, plus and minus the standard deviation is displayed in red; for VH the curves are displayed in cyan. Here $N = 200$, $\alpha = 0.05$. The number of coupons for RDS was set as 3.

3. The density of the graph was allowed to vary as in the previous experiment $\alpha \in \{0.01, 0.05, 0.1, 0.2\}$.

The parameters of the MRF are set up so $Q = 0.2$ for all regimes. Again, our method was implemented using the distributions described in Section 3.3; with the only modification that we used the model described in Expressions 8 and 9 as prior for the underlying network, instead of assuming an Erdős-Rényi model. Results are summarized in Table 3.

True $p(\mathcal{G})$	Bias (VH)	Cover (VH)	Length (VH)	Bias (Bayes)	Cover (Bayes)	Length (Bayes)
SW(2,0.95)	-0.0306	0.95	0.22	-0.0008	0.95	0.16
SW(10,0.95)	0.0001	0.95	0.25	-0.0009	0.98	0.12
SW(20,0.95)	0.001	0.95	0.25	-0.0008	0.96	0.13
SW(40,0.95)	-0.0082	0.9	0.25	-0.0009	0.95	0.13
SW(2,0.5)	-0.0204	0.95	0.25	-0.0006	0.95	0.14
SW(10,0.5)	-0.0173	0.96	0.26	-0.0001	0.96	0.13
SW(20,0.5)	-0.0138	0.95	0.24	-0.0009	0.99	0.12
SW(40,0.5)	-0.0092	0.93	0.26	0.0007	0.95	0.12
SW(2,0.25)	-0.0189	0.97	0.24	-0.0007	0.89	0.13
SW(10,0.25)	0.0104	0.91	0.24	0.0007	0.95	0.12
SW(20,0.25)	0.0158	0.94	0.25	0.0004	0.97	0.13
SW(40,0.25)	0.021	0.95	0.27	0.0006	0.95	0.13
SW(2,0.1)	-0.0173	0.9	0.23	-0.0002	0.63	0.13
SW(10,0.1)	-0.0255	0.9	0.26	-0.0007	0.75	0.13
SW(20,0.1)	-0.001	0.9	0.25	-0.0003	0.84	0.12
SW(40,0.1)	-0.001	0.9	0.26	0.0008	0.95	0.12

Table 3: Average bias, $|Q - \hat{Q}|$, and frequency of coverage for the Bayesian and non-model based approach. For the Bayesian method, the point estimator is given by the posterior mean and the 95% credible region was used for interval estimation. We compared these summaries to the Volz-Heckertorn (VH) estimator and the 95% bootstrap confidence interval (B-VH). The simulation scenarios are given by: the random graph model used to generate the data (Small-World in a circle), the average degree of the network, the probability of re-wiring and random graph model assumed when fitting the model (Erdős-Rényi). 100 simulations were performed for each scenario. In all scenarios the size of the network was set as 200. The average degrees were set in such way that they imply the density values 0.01, 0.05, 0.1 and 0.2. For each simulation, the BMA was implemented using 5 samples from the mixing distribution; for each of these samples, an MCMC was run using 3, 500 for burn-in and 500 posterior samples.

6 Real Data

We applied our methodology to the data derived from the study discussed in [de Mello et al. \(2008\)](#). This was a large RDS study implemented in a single location, namely the community of Campinas in the state of Sao Paulo, Brazil. Since RDS was used, non-ignorability is an issue for likelihood-based inferences. The aim of the study was to infer the prevalence of HIV among gay men in Campinas, Brazil.

The study comprised 658 men who have sex with men. The inclusion criteria used for this study, were:

1. born male;
2. had anal or oral sex with another man or transvestite in the past six months;
3. 14 years of age or older;
4. reside in the Metropolitan area of Campinas.

RDS was implemented using 16 seeds and a maximum of 3 referrals per subject (*i.e.*, $m = 3$). Point estimates (sample proportion and Volz-Heckathorn) and Bootstrap confidence intervals are shown in Table 4. The results shown in this table are not model-based.

	Naive		Volz-Heckathorn	
\hat{Q}	0.0789	(0.0577, 0.1001)	0.0711	(0.0466, 0.0955)

Table 4: Point estimators (sample proportion, Volz-Heckathorn) and the corresponding 95 per cent Bootstrap confidence intervals.

We also applied our method to this data set. We used the distributions described in Section 3.3. Since no prior information for the social network \mathcal{G} is available, a sensitivity analysis was conducted. For the Erdős-Rényi model, the density and the size of the graph were allowed to vary. Results from the sensitivity analysis are summarized in Table 5. Inferences tend to be quite stable with respect to the density and the network size assumed for the prior. Results differ from not model-based approaches; we claim that this is because we are correcting for biases due to the non-ignorability of the design. In addition it is not clear that the assumptions that guarantee the asymptotic unbiasedness of the Volz-Heckathorn estimator are met (see [Salganik and Heckathorn \(2004\)](#) and [Heckathorn \(2007\)](#) and the Section 7).

6.1 Assessing Goodness of Fit

To assess the goodness of fit of our model we made use of posterior predictive checks (See [Gelman et al. \(1996\)](#) and the appendix), *i.e.*, we selected a summary of the observed data and plotted the observed value of this summary against the distribution of replicates of such summary under the posterior predictive distribution. For this case, we used the sample mean of the observed Y 's as the summary. Results are displayed in Figure 7. The plot does not show evidence against the goodness of fit of our model.

N	Density	Mean	SD	0.025	0.05	0.5	0.95	0.97
1316	0.1	0.111	0.009	0.094	0.099	0.111	0.121	0.124
1316	0.05	0.112	0.009	0.095	0.098	0.112	0.126	0.130
1316	0.01	0.111	0.014	0.074	0.075	0.112	0.134	0.139
1316	0.005	0.117	0.012	0.102	0.104	0.117	0.132	0.135
1316	0.001	0.115	0.008	0.097	0.099	0.114	0.131	0.135
1316	$\frac{1}{N}$	0.112	0.009	0.099	0.101	0.112	0.123	0.128
2632	0.1	0.122	0.009	0.117	0.120	0.122	0.142	0.146
2632	0.05	0.124	0.015	0.116	0.120	0.124	0.148	0.150
2632	0.01	0.126	0.014	0.106	0.108	0.126	0.156	0.159
2632	0.005	0.128	0.012	0.120	0.121	0.129	0.164	0.168
2632	0.001	0.131	0.011	0.136	0.138	0.131	0.167	0.171
2632	$\frac{1}{N}$	0.127	0.012	0.094	0.099	0.127	0.150	0.155

Table 5: Summaries of the posterior for Q . These include: mean, standard deviation, and quantiles. Posterior samples were obtained assuming different values of the density for the Erdős-Rényi model and size of the underlying network N .

7 Discussion

In this paper we developed methodology for performing inference on non-ignorable designs on a network. The authors in [Handcock and Gile \(2011\)](#) discuss the idea of *amenable* designs and work with sampling mechanisms that fulfill that condition. It would be interesting to understand the relationship between amenability and graph ignorability. It is reasonable to think that the methodology proposed here can be used for dealing with designs that are not amenable. It is our understanding that all the available literature falls into one of two categories: Either they do not discuss ignorability, but assume implicitly that the sampling mechanism of their choice is ignorable (e.g., [Gile \(2011\)](#)), or they discuss ignorability and then they restrict their discussion to what they regard as ignorable designs [Handcock and Gile \(2011\)](#). In either case, the problem of making inference on a population quantity using a non-ignorable design is not addressed.

An important feature of the methodology we propose is that it is highly modular. By this we mean that the term $p(I \mid \mathcal{G})$ does not have to correspond to RDS. All the arguments hold for any other design that is not ignorable. The choices we made for $p(\alpha)$ and $p(\mathcal{G} \mid \alpha)$ were based on considerations such as simplicity and computational convenience. In principle, nothing prevents the reader from using a different random graph specification. Specifying the term $p(Y \mid \mathcal{G}, \gamma)$ is more delicate: Using a different MRF model would imply substantial changes in the MCMC procedure. Moving away from the MRF assumption would be even more challenging, and therefore an interesting line of research.

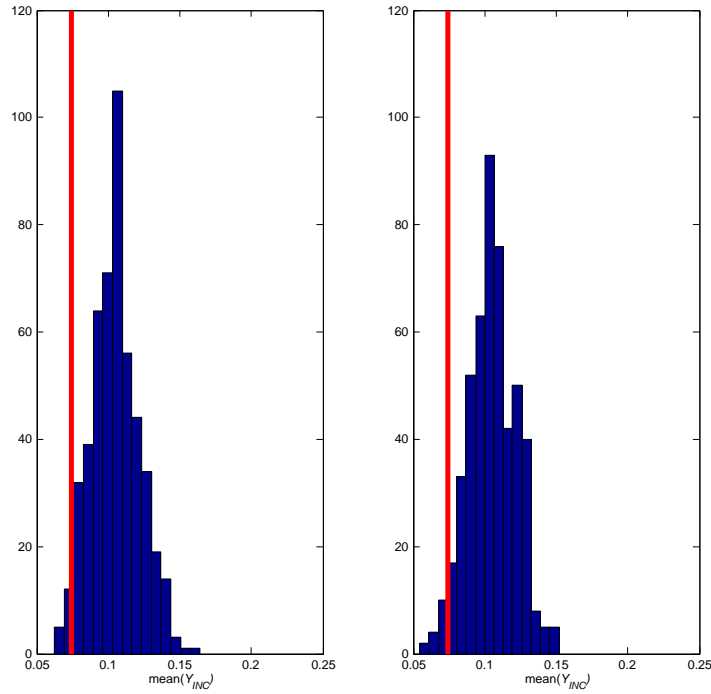


Figure 7: Sample mean of Y_{INC} (red line) compared to Monte Carlo distribution of repetitions of Y_{INC} obtained via the posterior predictive distribution. This was done for a networks with $N = 200$ and $\alpha = 0.01$ (left) and 0.05 (right). The same exercise was performed for $\alpha = 0.1$ (left) and 0.2 with similar results (not shown).

In, the authors present a set of assumptions that guarantee the asymptotic unbiasedness of the Volz-Heckathorn estimator ([Salganik and Heckathorn \(2004\)](#) and [Heckathorn \(2007\)](#)). These assumptions are:

1. If individual i can recruit individual j with positive probability, then the probability of j recruiting i must be positive, this is for all $1 \leq i < j \leq N$.
2. The graph \mathcal{G} is connected.
3. The proportion of sampled individuals is low enough so assuming that the sampling is with replacement is a reasonable approximation.

4. Respondents are able to accurately report their degree.
5. Each individual selects the subset of peers she (or he) will give the coupons according to an uniform distribution.

Assumptions 1 and 5 are phrased in terms of what is called the *respondent referral function* [Blitzstein and Nesterko \(2012\)](#). Note that our approach works for a general $p(I \mid \mathcal{G})$, therefore both assumptions become irrelevant for our framework. Since we can specify $p(\mathcal{G} \mid \alpha)p(\alpha)$ in our approach, we can deal with situations where there is not reasonable to assume \mathcal{G} to be connected (Assumption 2). Neither for the specification of the model or for the inference we need to assume sampling with replacement as an approximation for I , therefore our methodology is blind to Assumption 3. Our method could, in principle, incorporate information regarding the degree of the individuals included in the sample. If we wanted to do so, we could accommodate for a probabilistic model for the reported degree given the real degree. It follows that we could, in principle, deal with situations where Assumption 4 does not hold.

Our research suggests that RDS may not be the best way for collecting information about the parameters of the model, given all the sources of uncertainty. We are able to obtain good point estimates, but the associate uncertainty is higher than what most practitioners would like to afford. Future work includes taking the data generative process as a starting point (*i.e.*, the terms $p(\mathcal{G} \mid \alpha)$ and $p(Y \mid \mathcal{G}, \gamma)$) to design sampling mechanisms of the form $p(I \mid \mathcal{G})$ that would help to provide better inferences. That would imply establishing criteria for comparing sampling designs on networks, a remarkably unexplored area.

References

- Joseph Blitzstein and Sergiy Nesterko. Bias-variance and breath-depth tradeoffs in respondent-driven sampling. 2012.
- D. Brook. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51:481–483, 1964.
- M. de Mello, A.A. Pinho, M. Chinaglia, W. Tun, A. Barbosa Junior, and M.C.F.J. Ilario. Assessment of risk factors for hiv infection among men who have sex with men in the metropolitan area of campinas city, brazil, using respondent-driven sampling. Technical report, Washington DC: Population Council, 2008.

- Paul Erdos and A. Renyi. The evolution of random graphs. *Magyar Tud. Akad. Mat. Kutato Int. Kolz*, 5:17–61, 1960.
- Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.
- Krista J. Gile. Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *Journal of the American Statistical Association*, 106:135–146, 2011.
- Mark S. Handcock and Krista J. Gile. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4:5–25, 2011.
- Douglas D. Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44:174–199, 1997.
- Douglas D. Heckathorn. Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential degree. *Sociological Methodology*, 37:151–207, 2007.
- D.F. Heitjan and Donald B. Rubin. Ignorability and coarse data. *Annals of Statistics*, 19:2244–2253, 1991.
- Mark S. Kaiser and Noel Cressie. The construction of multivariate distributions from markov random fields. *Journal of Multivariate Analysis*, 73:199–220, 2000.
- Jun Liu and R. Chen. Blind deconvolution via sequential imputations. *Journal of the American Statistical Association*, 90:567–576, 1995.
- Patrick Perry and Patrick Wolfe.
- A. Raftery, D. Madigan, and C. Volinsky. Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). volume 5 of *Bayesian Statistics*, pages 323–349. Oxford University Press, 1996.
- Christian P. Robert. *The Bayesian Choice, Second Edition*. Springer-Verlag, 2001.
- Christian P. Robert and George Casella. *Monte Carlo Statistical Methods, Second Edition*. Springer-Verlag, 2004.
- Donald B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.
- Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, 1987.

Matthew J. Salganik and Douglas D. Heckathorn. Sampling and estimation in hidden populations using respondent driven sampling. *Sociological Methodology*, 34:193–239, 2004.

Erik Volz and Douglas D. Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24:79–97, 2008.

A Appendix: Computing Joint Density for a MRF

We now describe briefly the algorithm proposed by Kaiser and Cressie [Kaiser and Cressie \(2000\)](#) for computing the joint distribution of MRF from the full conditionals. A key concept is the one of *negpotential function*:

$$Q(y) \equiv \log \left\{ \frac{p(y \mid \mathcal{G}, \gamma)}{p(y^* \mid \mathcal{G}, \gamma)} \right\}, \quad y \in \Omega, \quad (17)$$

where $y^* \in \Omega$ is such that $p(y^* \mid \mathcal{G}, \gamma)$ is finite. $Q(y)$ can be expanded in the following way:

$$\begin{aligned} Q(y) &= \sum_{1 \leq i \leq n} H_i(y_i) + \sum \sum H_{i,j}(y_i, y_j) \\ &\quad + \sum \sum \sum H_{i,j,k}(y_i, y_j, y_k) \\ &\quad \dots \\ &\quad + H_{1,2,\dots,n}(y_1, y_2, \dots, y_n), \quad y \in \Omega. \end{aligned}$$

The usefulness of this fact is that each of the H functions can be expressed in terms of the full conditionals:

$$\begin{aligned} H_{i,j(2),\dots,j(m)}(y_i, y_{j(2)}, \dots, y_{j(m)}) \\ &= \sum_{t=0}^{m-2} \sum_{j_{m-t} \in T_m(m-t)} \\ &\quad \times \left\{ (-1)^{t-1} \log \left[\frac{p(y_i \mid \{y_k : k \in j_{m-t}^i\}, \{y_k^* : k \notin j_{m-t}\}, \mathcal{G}, \gamma)}{p(y_i^* \mid \{y_k : k \in j_{m-t}^i\}, \{y_k^* : k \notin j_{m-t}\}, \mathcal{G}, \gamma)} \right] \right\} \\ &\quad + (-1)^{m-1} \log \left[\frac{p(y_i \mid \{y_j^* : j \neq i\}, \mathcal{G}, \gamma)}{p(y_i^* \mid \{y_j^* : j \neq i\}, \mathcal{G}, \gamma)} \right] \end{aligned}$$

Clearly if

$$y_i = y_i^*, y_{j(2)} = y_{j(2)}^*, \dots, y_{j(m)} = y_{j(m)}^*,$$

then $H_{i,j(2),\dots,j(m)} = 0$. In addition, according to Theorem 2 in [Kaiser and Cressie \(2000\)](#), any H function is equal to zero unless the corresponding vertices in \mathcal{G} form a clique. Theorem 3 in [Kaiser and Cressie \(2000\)](#) states that, under certain conditions, the joint $p(y \mid \mathcal{G}, \gamma)$ can be computed by first computing the sum:

$$\sum_{\omega \in \Omega} \exp \{Q(y(\omega))\},$$

and then

$$p(y \mid \mathcal{G}, \gamma) = \frac{\exp \{Q(y)\}}{\sum_{\omega \in \Omega} \exp \{Q(y(\omega))\}}.$$

B More on Simulations

Density	N	m	Bias Q_∞	Coverage Q_∞	Bias Q_{Emp}	Coverage Q_{Emp}	Length	Method
0.01	100	3	-0.0096	0.91	-0.0012	0.93	0.11	Bayes
0.01	100	3	0.0105	0.92	-0.0129	0.98	0.25	VH
0.05	100	3	-0.0007	0.93	0.0030	0.94	0.11	Bayes
0.05	100	3	-0.0004	0.91	0.0330	0.96	0.25	VH
0.1	100	3	0.0003	0.93	0.0006	0.97	0.11	Bayes
0.1	100	3	0.0084	0.92	0.0087	0.92	0.25	VH
0.2	100	3	-0.0051	0.94	-0.0024	0.95	0.11	Bayes
0.2	100	3	-0.0067	0.94	-0.0040	0.91	0.26	VH
0.01	100	5	-0.0001	0.9	-0.0033	0.91	0.11	Bayes
0.01	100	5	0.0018	0.92	-0.0014	0.94	0.24	VH
0.05	100	5	0.0005	0.94	0.0008	0.93	0.11	Bayes
0.05	100	5	-0.0018	0.96	-0.0015	0.91	0.27	VH
0.1	100	5	-0.0026	0.93	0.0010	0.96	0.11	Bayes
0.1	100	5	-0.0123	0.93	-0.0087	0.93	0.29	VH
0.2	100	5	0.0057	0.99	0.0048	0.95	0.11	Bayes
0.2	100	5	0.0039	0.89	0.0030	0.93	0.29	VH
Density	N	m	Bias Q_∞	Coverage Q_∞	Bias Q_{Emp}	Coverage Q_{Emp}	Length	Method
0.01	200	3	-0.0062	0.93	-0.0052	0.93	0.12	Bayes
0.01	200	3	0.0017	0.89	0.0027	0.92	0.24	VH
0.05	200	3	-0.0043	0.92	-0.0001	0.93	0.11	Bayes
0.05	200	3	-0.0095	0.91	-0.0053	0.95	0.26	VH
0.1	200	3	0.0035	0.93	0.0073	0.91	0.11	Bayes
0.1	200	3	0.0018	0.90	0.0056	0.89	0.25	VH
0.2	200	3	0.0006	0.98	-0.0017	0.89	0.10	Bayes
0.2	200	3	0.0077	0.92	0.0090	0.92	0.23	VH
0.01	200	5	0.0005	0.9	0.0020	0.93	0.11	Bayes
0.01	200	5	0.0040	0.89	0.0054	0.90	0.23	VH
0.05	200	5	0.0001	0.94	0.0080	0.92	0.11	Bayes
0.05	200	5	0.0011	0.84	0.0005	0.86	0.27	VH
0.1	200	5	-0.0008	0.96	0.0008	0.92	0.11	Bayes
0.1	200	5	-0.0003	0.89	0.0013	0.89	0.28	VH
0.2	200	5	-0.0007	0.98	-0.0019	0.94	0.11	Bayes
0.2	200	5	-0.0061	0.88	-0.0073	0.88	0.28	VH

Table 6: Average bias, $Q - \hat{Q}$, and frequency of coverage for the Bayesian and non-model based approach. For the Bayesian method, the point estimator \hat{Q}_B is given by the posterior mean and the 95% credible region was used for interval estimation. We compared these summaries to the Volz-Heckartorn (VH) estimator and the 95% bootstrap confidence interval. The simulation scenarios are given by: Density of the underlying network, and the maximum number of referrals m . The size of the underlying network was set as 200 and the sample size as 50. 100 simulations were performed for each scenario. In all cases an Erdős-Rényi model is used to generate the data. For each simulation, the BMA was implemented using 5 samples from the mixing distribution; for each of these samples, an MCMC was run using 3,000 for burn-in and 500 posterior samples.

Density	N	m	Bias Q_∞	Coverage Q_∞	Bias Q_{Emp}	Coverage Q_{Emp}	Length	Method
0.01	500	3	0.0009	0.93	0.0013	0.93	0.10	Bayes
0.01	500	3	-0.0028	0.93	-0.0024	0.95	0.25	VH
0.05	500	3	0.0060	0.92	0.0330	0.92	0.10	Bayes
0.05	500	3	0.0036	0.94	0.0009	0.94	0.26	VH
0.1	500	3	0.0015	0.92	0.0038	0.90	0.09	Bayes
0.1	500	3	0.0011	0.94	0.0340	0.95	0.26	VH
0.2	500	3	-0.0065	0.91	-0.0077	0.91	0.08	Bayes
0.2	500	3	-0.0122	0.94	-0.0134	0.95	0.27	VH
0.01	500	5	-0.0004	0.91	-0.0010	0.91	0.10	Bayes
0.01	500	5	0.0060	0.93	0.0054	0.93	0.27	VH
0.05	500	5	0.0030	0.93	0.0019	0.89	0.09	Bayes
0.05	500	5	0.0023	0.93	0.0013	0.92	0.28	VH
0.1	500	5	0.0036	0.94	0.0029	0.92	0.10	Bayes
0.1	500	5	0.0027	0.91	0.0020	0.90	0.28	VH
0.2	500	5	-0.0014	0.95	0.0039	0.94	0.09	Bayes
0.2	500	5	-0.0011	0.83	-0.0001	0.83	0.27	VH

Table 7: Average bias, $Q - \hat{Q}$, and frequency of coverage for the Bayesian and non-model based approach. For the Bayesian method, the point estimator \hat{Q}_B is given by the posterior mean and the 95% credible region was used for interval estimation. We compared these summaries to the Volz-Heckatortn (VH) estimator and the 95% bootstrap confidence interval. The simulation scenarios are given by: Density of the underlying network, and the maximum number of referrals m . The size of the underlying network was set as 200 and the sample size as 50. 100 simulations were performed for each scenario. In all cases an Erdős-Rényi model is used to generate the data. For each simulation, the BMA was implemented using 5 samples from the mixing distribution; for each of these samples, an MCMC was run using 3,000 for burn-in and 500 posterior samples.